

**ISSS617: Python for Data Science**

**PROJECT FINAL REPORT**

**Group G1-6**

Akash Khowala

Amit Singla

Neel Chomal

Pankaj Bhootra

Wu Yinxiao

**Topic:** Trump Twitter Analysis

**Jupyter Notebook:** <https://github.com/akhowala/TwitterAnalysis>



*We understand what plagiarism is and have ensured we did not plagiarise for this assignment. This assignment is in partial fulfilment of the requirements for the module ISSS617 Python for Data Science.*

**TABLE OF CONTENTS**

**1. Introduction 3**

**2. Scope of Problem and Stakeholders 4**

**3. Previous Work 4**

**4. Datasets Used 6**

**5. Implementation 7**

5.1 Step-by-Step 7

5.2 Libraries Used 7

**6. Data Pre-processing 8**

**7. Topic Modelling 9**

7.1 Introduction 9

7.2 Approach 9

7.3 Results 9

**8. Sentiment Analysis 13**

8.1 Introduction 13

8.2 Approach 14

8.3 Results 15

**9. Descriptive Data Analysis 16**

**10. Predictive Modelling 18**

**11. Future Work 20**

**12. Conclusion 20**

**1. Introduction**

Our objective is to explore the idea whether powerful and influential people’s statements on social media have an impact on financial markets. We have decided to study the tweets of Donald Trump, President of the United States, and study his impact on US Treasury markets. Since 2009, Trump has been actively tweeting and expressing his opinions on global issues such as politics, military, markets, trade, economy, immigration, etc.

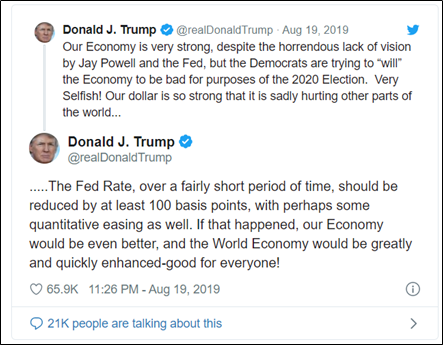
Research has shown that social media plays a leading role in forecasting events, using thoughts of people by collecting, storing and analysing data for the purpose of harvesting information. Social media analytics is focused on developing frameworks to collect, monitor, analyse, summarize, and visualize social media data to extract useful patterns. We will analyse all the tweets posted by Trump through various Twitter handles (@POTUS, @realDonaldTrump, and relevant allies) to understand the topics he posts about, the sentiments expressed in the tweet (positive, negative, neutral), run a timeline analysis by comparing the sentiments during a period with the trends in US treasury market during the same period and build a predictive model for the same. We have a large corpus of Donald Trump tweets going back several years, along with data on historical market trends, which will give us enough insight to capture his impact and draw relevant conclusions.

Figure 1: An example of Trump's tweet which could possibly affect the markets

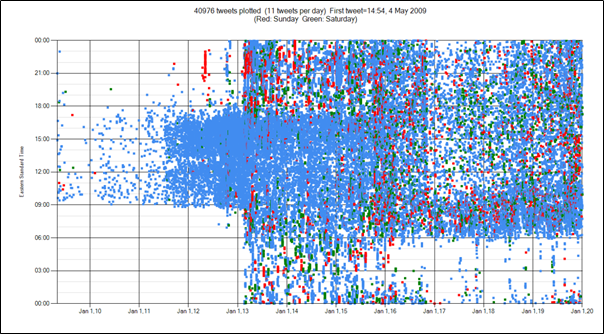


Figure 2: Donald Trump's tweet activity from his first tweet in May '09. His tweet activity pattern has changed since 2013

**2. Scope of Problem and Stakeholders**

This project will only cover Trump’s potential impact on US Treasury markets and as such, we will run topic modelling and sentiment analysis algorithms to filter out irrelevant tweets. A potential application of our project could be around building a predictive market index to capture the market impact of Trump’s tweets in real-time, for which the stakeholders would include banks, brokerages and investors, who are interested in predicting market trends accurately.

**3. Previous Work**

Major players in the financial markets have tried to infer market trends from Trump’s tweets. JP Morgan’s analysts have created the “Volfefe Index”. For every 5 minutes after Trump tweeted with direct Fed references, the index showed “a rolling one-month probability” that each missive is market moving for treasury yields.

Citi has similarly shown that Trump’s tweets are generally followed by volatile behaviour across currency markets globally. They have forecasted that markets are likely to show more sensitivity to Trump’s continued tweeting with USD and Fed references.



Figure 3: JP Morgan created a 'Volfefe Index', which shows “a rolling one-month probability” that each tweet is market moving

In another study, analysts at Bank of America Merrill Lynch published a note concluding that days during which Trump tweets relatively frequently tend to see negative stock returns of 9 basis points on average. Days with fewer presidential tweets tend to see positive returns of 5 basis points on average. The S&P 500 is up more than 35% since Trump won the 2016 election.

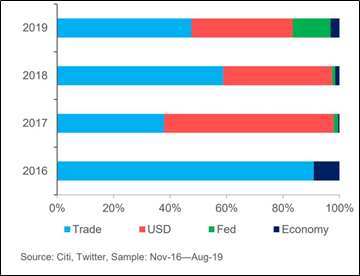
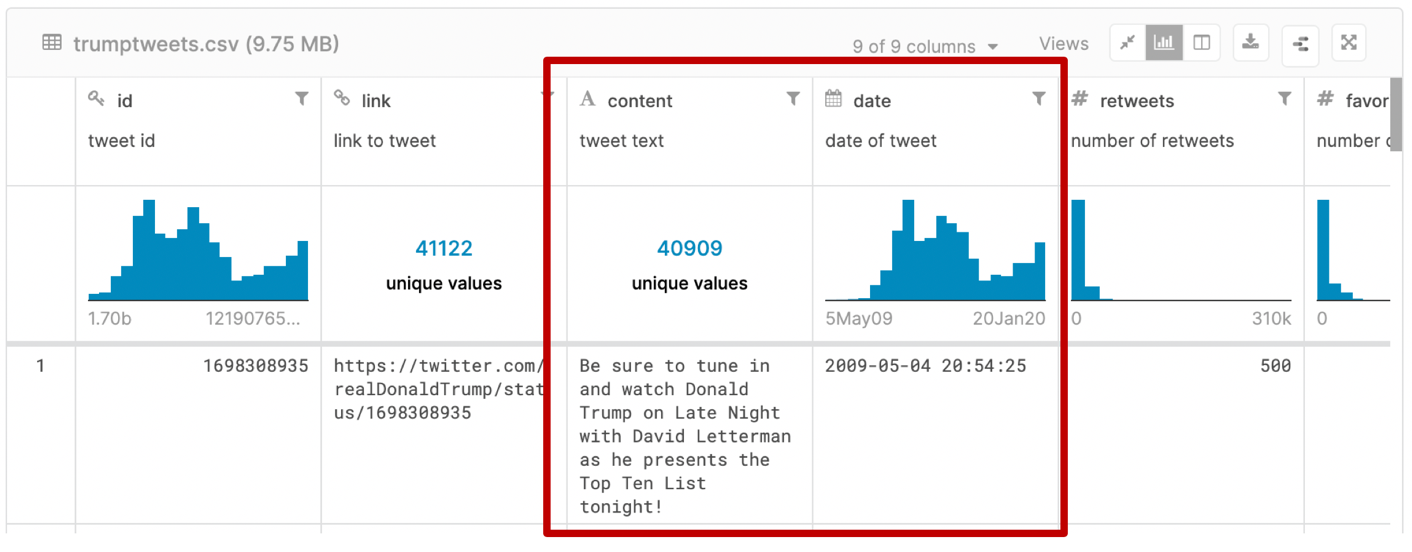


Figure 4: Major market moving topics which Trump has tweeted about since taking over as President (ref. Citi)

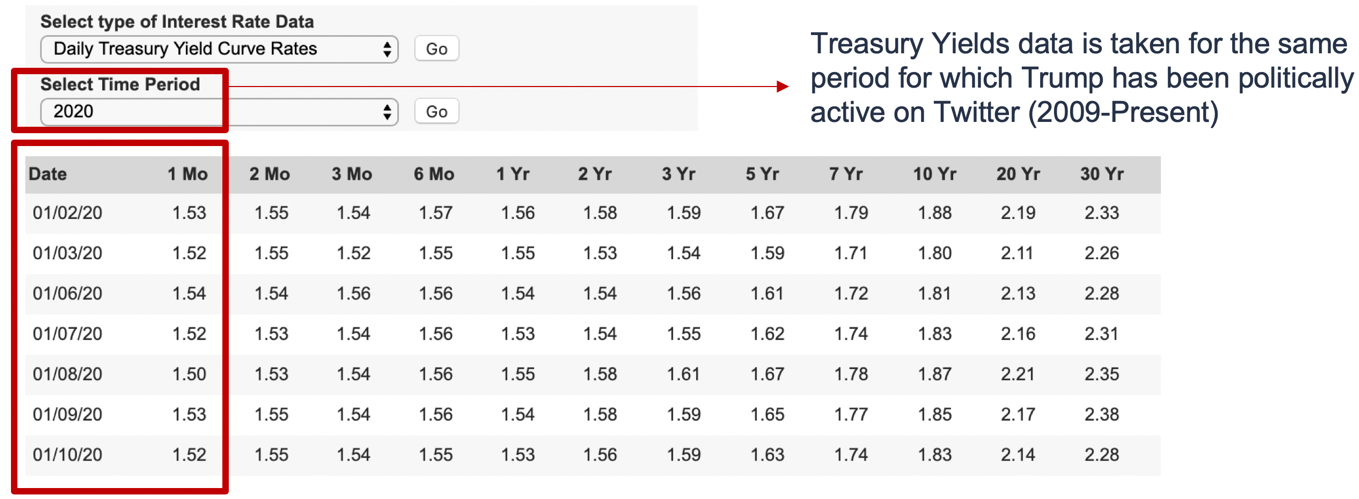
**4. Datasets Used**

1. [Kaggle dataset of Trump Tweets](https://www.kaggle.com/austinreese/trump-tweets)



This dataset contains a list of all tweets from Donald Trump since he has been politically active on Twitter (2009 and after). For every tweet, we are interested in the content (for Sentiment Analysis and Topic Modelling) and the date (for EDA with market data).

2. [US Treasury Yield Curve Rates](https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield)



For US treasury market, this provides the historical yield curve rates, out of which we are interested in the short term (1 month) projections only as we expect that Trump’s tweets will only, potentially, have short term impact on the treasury market.

**5. Implementation**

**Goal:** Use Trump’s sentiments on particular topics to estimate the short term market movement.

**5.1 Step-by-Step**

1. **Data Pre-processing:** We start with some clean-up of tweets dataset and vectorize it, along with some pre-computations for the treasury yields data.
2. **Topic Modelling:** We take the vectorized tweets and run a clustering algorithm named Latent Dirichlet Allocation (LDA) to break the tweets into clusters, each corresponding to a unique topic.
3. **Sentiment Analysis:** We take the pre-processed tweets data (not vectorized) and use existing libraries to estimate a sentiment polarity score for each tweet.
4. **Descriptive Data Analysis:** We need to filter out topics/sentiments which do not have any impact on the market. Hence, for every combination of topic/sentiment, we generate visualizations highlighting its impact on the market movement (up/down/neutral).
5. **Predictive Modelling:** Training a logistic regression model on only market impacting tweets and optimizing its accuracy for future estimation of treasury market movement.

**5.2 Libraries Used**

To implement the above approach, we have used the following libraries across different tasks:

1. **Data Manipulation:** numpy, pandas
2. **Data Pre-processing:** nltk, re, wordcloud
3. **Topic Modelling:** sklearn, pyLDAvis, pickle
4. **Sentiment Analysis:** TextBlob, vaderSentiment
5. **Exploratory Data Analysis:** matplotlib, seaborn

**6. Data Pre-processing**



Figure 5: Word Cloud for Pre-processed Tweets

1. Pre-processing for Tweets (using regex and nltk)

* Removing tweet hyperlinks and usernames (words starting with @), hashtags are retained
* Removing punctuations (any non-alphanumeric character)
* Converting to lowercase
* Removing stopwords
* Word Net Lemmatization, converting inflected forms of a word to the base word

Tweets are further vectorized using the CountVectorizer from sklearn (Bag of Words approach).

2. Pre-processing for Treasury Yields dataset (using numpy)

The raw dataset gives absolute daily figures but not the difference from previous days’ figures. This daily difference is calculated and mapped to an “UP”, “DOWN” or “NEUTRAL” label. For a tweet made before 3:00 PM, we have compared the treasure yield price of the same day and compared with the price on the previous day. Otherwise, for tweets made after 3:00 PM till 11:59 PM, we look at price for the next day and compare it with the price on the preceding day.

**7. Topic Modelling**

**7.1 Introduction**

Topic Modelling is a Natural Language Processing (NLP) technique that allows us to automatically extract meaning from texts by identifying recurrent themes or abstract “topics”. We are using Latent Dirichlet Allocation (LDA) to cluster Trump’s tweets into topics such as jobs, trade, military, immigration, etc. LDA allows sets of observations to be explained by ‘unobserved’ groups, in this case words collected into tweets, where each tweet is a mixture of one or more topics and that each word’s presence is attribute to one of the tweet’s topics.

**7.2 Approach**

Using the vectorized tweets dataset, we use sklearn’s API to train and fit LDA models on varying number of topics, from 10 to 250 and for each model, the perplexity score is calculated. To further validate the quality of clusters obtained in each trained model, we use the library pyLDAvis to interpret the topic clusters in the form of an interactive web-based visualization, which shows a relative distribution of topics, or an inter-topic distance map, in the 2D coordinate space and for each cluster, the top 30 salient terms found in that cluster, which we use to label every cluster, and in turn every tweet, with a topic.

**7.3 Results**

Perplexity is a statistical measure of how well a probability model predicts a sample. We want to achieve a lower perplexity number, but we also want to be able to interpret the topics easily. We have trained 10 to 250 topics, and from the graph on next page, we conclude that topics between 10 to 200 has a lower perplexity level than over 200. (Figure 1 Perplexity graph for models of 10 to 250 topics). Between 10 to 200 topics, if topics is less than 30, the model becomes difficult to interpret and if over 75 topics are chosen, the tokens become too scattered with very small token size in each topic. Therefore, we will focus on 30-75 topics where perplexity score is relatively low and models are clearer for interpretation.

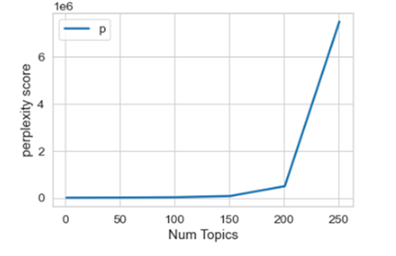


Figure 6: Perplexity graph for model of 10-250 topics

Trump has been President of the United States since January 2017. In 2016, Trump had 32 million followers, and now he has 76 million followers, more than double before he assumed presidency. We have trained 3 LDA models based on Trump’s twitter dataset. The first model is before he assumed presidency with data timeframe between 2009 to 2016. The second model is focused on tweets from 2017 onwards after he became President. We will analyse his top 10 words before and then we will compare his major topics, both before and after assuming presidency. The third model is based on using only ‘volatile’ data. We define ‘volatile’ as days on which the market movement in treasury yield curve rate was either “UP” or “DOWN” (not “NEUTRAL”). There are about 32k tweets, but only about 4000 with ‘no neutral’ days.

In Figure 7, Top 10 words before/after Trump assumed Presidency, we can tell Trump mentioned “people”, “job” and “border” significantly more often than before. He mentioned “people” 140 times per year before 2017, and 266 time per year after 2017. He talked about “job” twice as much as before and started increasing the voice on border issue with Mexico.

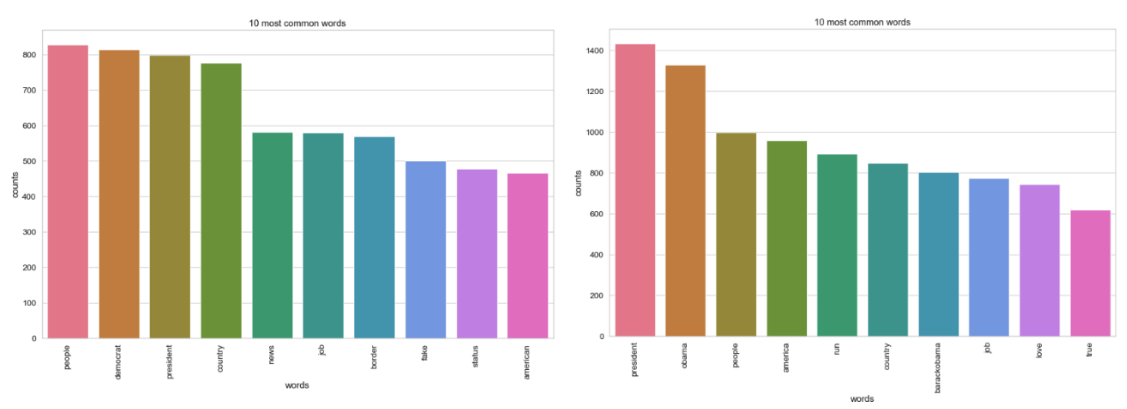
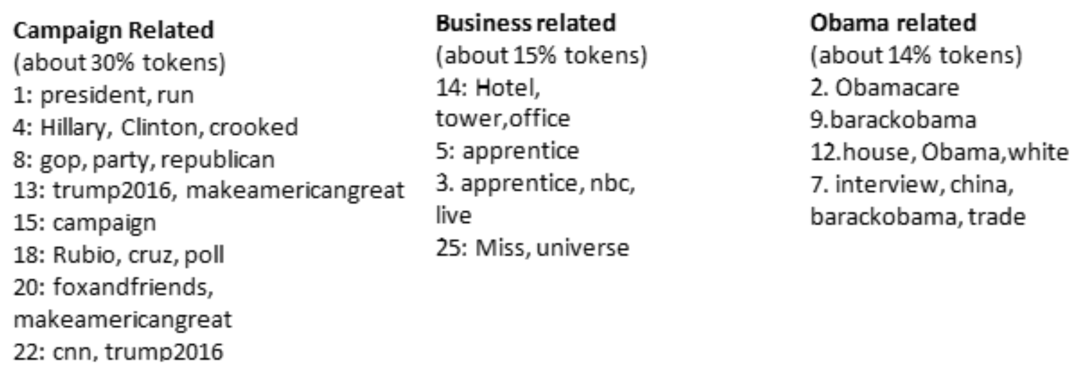


Figure 7: Top 10 words before and after Trump assumed Presidency

Trump has been politically active since 2009. Based on our first model, his major topics are “Campaign related”, “Business related” and “Obama related”. Most of his political activity is concentrated on campaigns. About 30% of the tokens are clearly related to presidency election in 2016. 15% of the token are about his own business, his TV show apprentice and Miss Universe which trump owned for decades.



After Trump has become president, his major topics have shifted from campaigns and private businesses to US politics, trade issues, national security and immigration. 20% of tokens post presidency are related to US political issues and 12% tokens are regarding topics of security and immigration. The most interesting group, however, is the Trade Issue topic. This is mainly related to words like “china”, “trade”, “deal”, “fed”, “stock market” etc. This is a good example of a topic potentially correlated to volatility in treasury yield curve rates.



The last model is based on the volatile data during Trump’s presidency. We took out “flat” days when there was no movement in treasury yield rates, so only tweets for days with “UP” or “DOWN” market movement were retained. From the top 30 salient terms across topics, we can tell that all these topics are good candidates for impacting the treasury yield market rates.

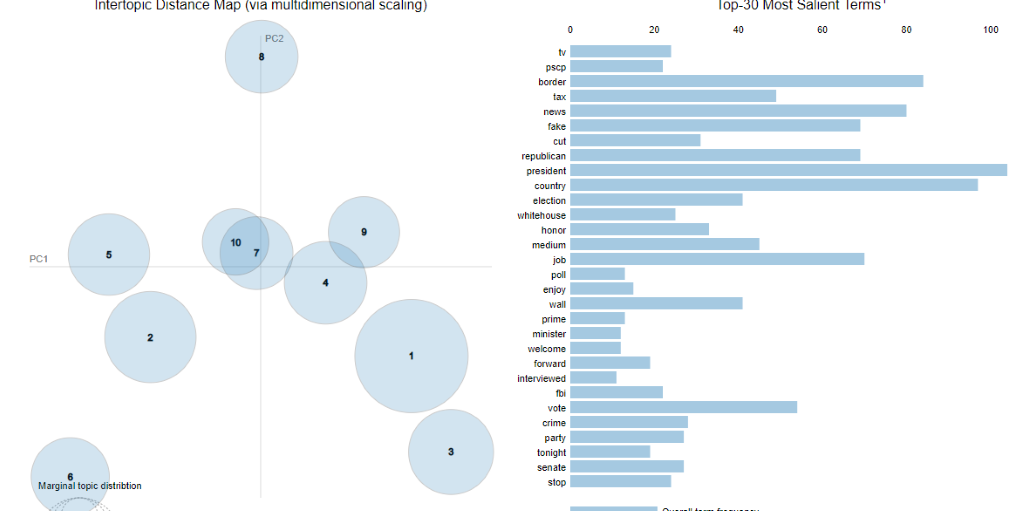


Figure 8: Visualization of Topics using pyLDAvis

**8. Sentiment Analysis**

**8.1 Introduction**

Sentiment refers to attitude, emotions, and opinions. Sentiment Analysis quantifies the tone of a piece of text by identifying and scoring positive and negative words. Sentiment Analysis is contextual mining of text which identifies and extracts subjective information in source material which helps a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This basic analysis can also prove to be very handy with further exploratory data analysis and deriving useful business meanings.

Below are the main types of sentiment analysis:

* Fine-grained Sentiment Analysis involves determining the polarity of the opinion. It can be a simple binary positive/negative sentiment differentiation. This type can also go into the higher specification (for example, very positive, positive, neutral, negative, very negative), depending on the use case (for example, as in five-star Amazon reviews). We have used this type of analysis in our project with focus on positive/negative/neutral tweets only.
* Emotion detection is used to identify signs of specific emotional states presented in the text. Usually, there is a combination of lexicons and machine learning algorithms that determine what is what and why.
* Aspect based Sentiment Analysis goes deeper. Its purpose is to identify an opinion regarding a specific element of the product. For example, the brightness of the flashlight in the smartphone. The aspect-based analysis is commonly used in product analytics to keep an eye on how the product is perceived and what are the strong and weak points from the customer point of view.
* Intent Analysis is all about the action. Its purpose is to determine what kind of intention is expressed in the message. It is commonly used in customer support systems to streamline the workflow.

**8.2 Approach**

We have primarily used TextBlob for our sentiment analysis. TextBlob is a Python library for processing textual data. It provides APIs for common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

TextBlob can work with different machine learning models used in natural language processing. We create a classifier to analyse the polarity of each tweet after cleaning the tweets. The way it works is that TextBlob already provides a trained analyser. Training data now consists of labelled positive and negative features. This data is trained on a Naive Bayes Classifier. Finally, parsed tweets are returned.

The sentiment function of TextBlob returns two properties, polarity and subjectivity.

* Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.
* Subjectivity is also a float which lies in the range of [0,1]. Subjective sentences (closer to 1) generally refer to personal opinion, emotion or judgment whereas objective (closer to 0) refers to factual information.

We have also validated our results using VADER to get a reasonably similar result. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative. VADER is fully open sourced and has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews.

**8.3 Results**

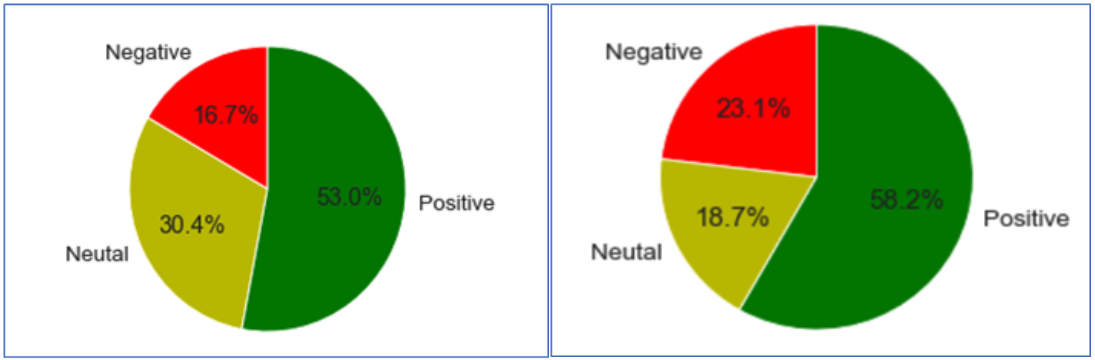


Figure 9: Results for Sentiment Analysis – TextBlob (left) and Vader (right)

With both TextBlob and Vader, sentiment analysis is done to find the percentage of positive, negative and neutral tweets and visualize on the pie charts as shown above. Analysing the same set of data using two libraries here and randomly checking for the classification for the same testing set of tweets, we validated that the results from both libraries are reasonably consistent and similar.

Sentiment Analysis is a great way to understand the general opinion of the public. In this case, the focus has been on Trump’s tweets and we have been able to perform further analysis using the sentiments involved in his tweets. However, it has its own set of challenges and limitations. Sometimes, it is difficult to understand the tone of a tweet, especially if there is irony or sarcasm involved, which remains a general issue for further research and refinement.

**9. Descriptive Data Analysis**

After we have complete dataset including Tweets (text, clustered under topics, and sentiments) along with up/down movement indicator for “One Monthly return” on Treasury prices, we explore the dataset for interesting trends.

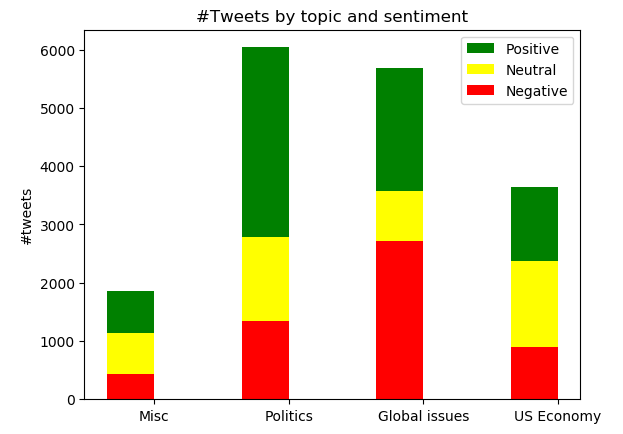


Figure 10: Clustered tweets and sentiment split

Majorly the tweets were around Politics (US elections, Hilary, Obama, Obamacare, etc.) followed by “Global issues” like trade with China, Mexico border, Iran and many others. The sentiments were majorly positive apart from negative connotation on Global issues.

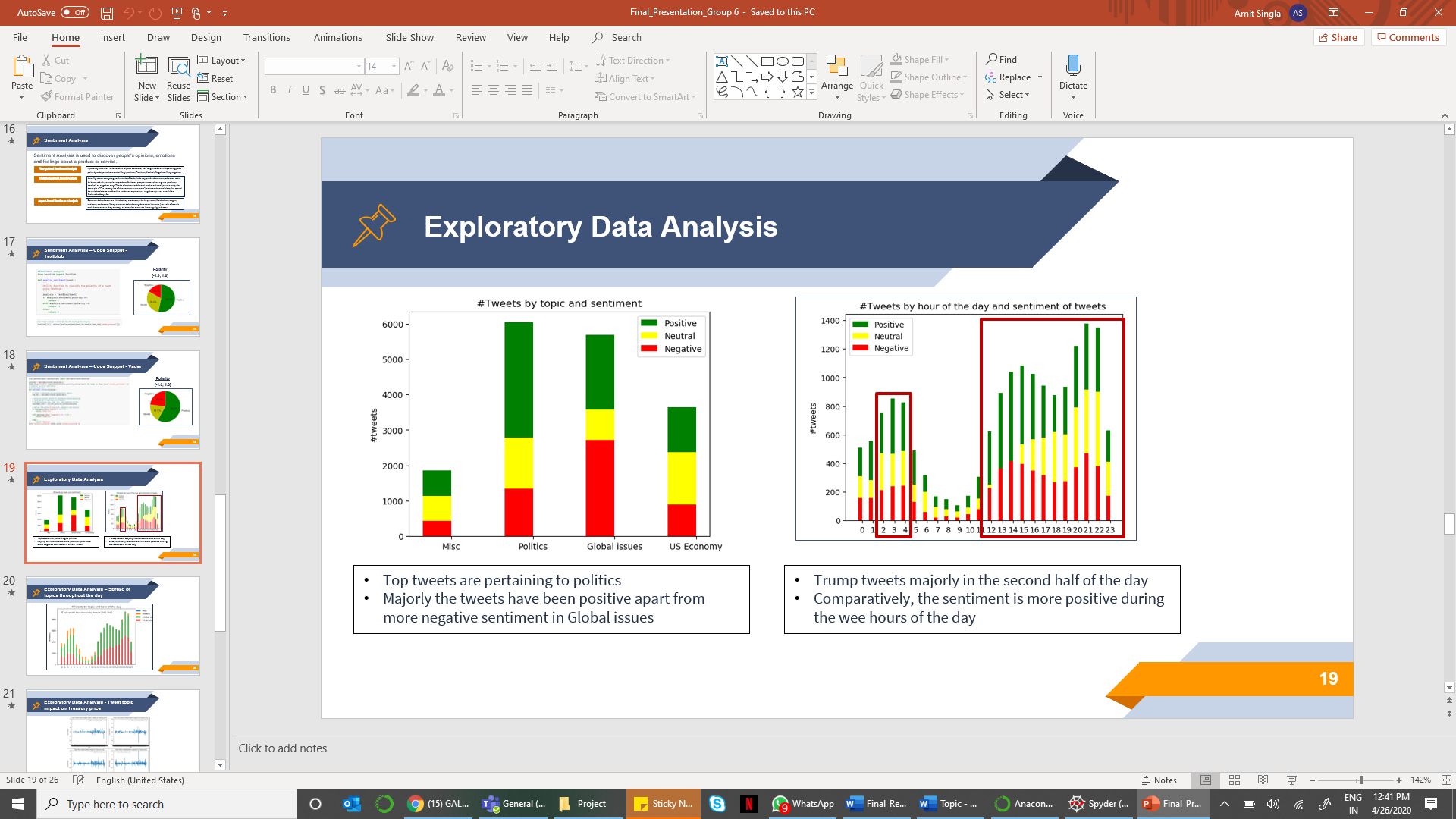


Figure 11: Spread of Sentiment throughout the day

The average #tweets/hour were high in second half of the day and some in wee hours of the day (between 2:00- 4:00 AM). Sentiments are highly negative around afternoon.

A picture containing pencil

Description generated with very high confidence

Figure 12: Spread of topics during the day

Tweets about the different topics were spread throughout the day as we can find from the chart above.

A screenshot of a cell phone

Description generated with very high confidence

Figure 13: Impact of selected topics on treasury rates

Further, we analyse the data market response in connect with different topics Trump tweeted about along with taking into consideration the sentiment of the tweet. As visible from the chart 4 about Politics, we see major churn in prices.

A screenshot of a social media post

Description generated with very high confidence

Figure 14: Topic sentiment impact on price movement

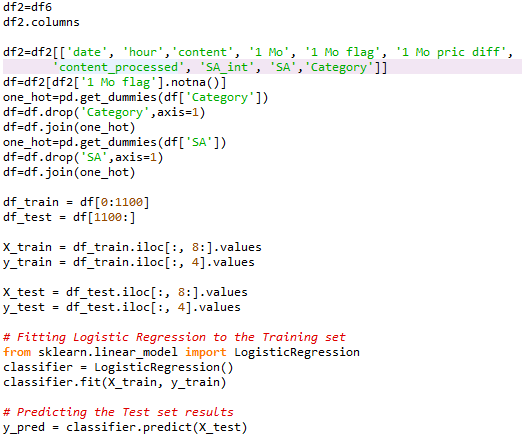
Charts in top row are price fluctuation due to neutral tweets, while charts in middle row are movement in price due to positive perceived tweets and last row charts are for negative tweets impact on treasury price movement. Clearly, comparatively very low-price movement for neutral tweets while amplitude of treasury price churn is high for positive and negative tweets.

While there were many interesting trends emerging around the descriptive analysis, however we were looking for combination of prominent topics and sentiments contributing to price fluctuation. However, it did not emerge out from the dataset prepared.

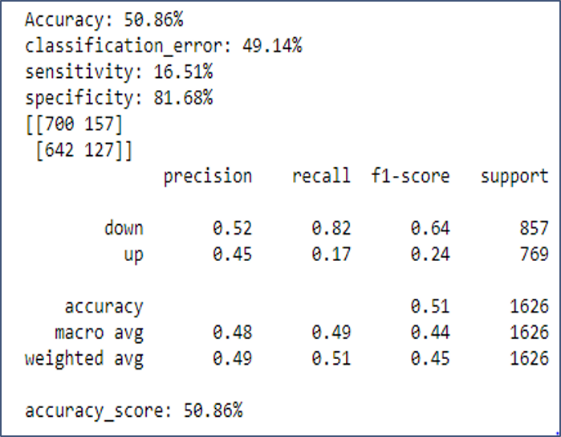
**10. Predictive Modelling**

As shared, no dominant topic/sentiment seems to stand out to explain price fluctuations. Hence, we have used all topics and sentiments while creating a predictive model using Linear Regression module. However, we have filtered out the dataset where the price fluctuation in comparison to the previous day was neutral. Hence, this leaves with comparatively much lesser however still enough number of rows (around 4k).

Below is a snapshot of the code used to train the model. We have equally split the dataset into train and test. Hot encoded the columns for sentiment and topics and dropped the root columns to avoid redundancy.



To measure the accuracy of the model, we have computed confusion matrix, sensitivity, specificity, and misclassification rate.



It is evident from the KPI scores that the model requires a lot of fine tuning due to high misclassification rate at 49% and very low sensitivity scores.

**11. Future Work**

1. Advanced Machine Learning – We feel that there is potential to improve the performance of every step in our pipeline, especially on pre-processing, vectorizing, topic modelling and sentiment analysis. We hope to learn advanced ML in future courses and apply the knowledge to achieve better end results in our predictive modelling.
2. Other financial markets – Per the exploratory data analysis, it was found that US treasury yield curve rates tend to be more stable in comparison to equity markets. Hence, we will extend our efforts to analyse the impact of POTUS tweets on other financial markets.
3. Build the next “Volfefe Index” – JP Morgan built an ambitious market index to analyse real-time impact of Trump’s tweets. With accurate classification in predictive modelling, we hope to extend this for other markets.

**12. Conclusion**

With Donald Trump’s tweets, we have a splendid example of a social analytics use case that can potentially provide accurate estimation for market trends. We have built an end-to-end pipeline to investigate a correlation between Trump’s sentiments on certain topics and the corresponding movement of the US treasury rates. We conducted an extensive Descriptive Data Analysis and trained classification models on tweets from all combinations of topics/sentiments to optimize the accuracy. After thorough experimentation, we found that US treasury yield rates tend to differ by no more than 30 cents for a single day movement as a result of Trump’s tweets, which indicates that this market is mostly stable. Also, the solution we have developed is easily extensible to other influential figures and the corresponding, susceptible financial markets. We are excited about this problem domain and look forward to discovering more patterns in future endeavours.